Genetics: Early Online, published on July 31, 2018 as 10.1534/genetics.118.300733

Geometry of the sample frequency spectrum and the perils of demographic inference

Zvi Rosen[*,1], **Anand Bhaskar**[†,‡,1], **Sebastien Roch**[§] **and Yun S. Song**[*,**,††,2]

[*]Department of Statistics, University of California, Berkeley, CA 94720, USA, [†]Department of Genetics, Stanford University, Stanford, CA 94305, USA, [‡]Howard Hughes Medical Institute, Stanford University, Stanford, CA 94305, USA, [§]Department of Mathematics, University of Wisconsin, Madison, WI 53706, USA, [**]Computer Science Division, University of California, Berkeley, CA 94720, USA, [††]Chan Zuckerberg Biohub, San Francisco, CA 94158, USA

ABSTRACT The sample frequency spectrum (SFS), which describes the distribution of mutant alleles in a sample of DNA sequences, is a widely-used summary statistic in population genetics. The expected SFS has a strong dependence on the historical population demography and this property is exploited by popular statistical methods to infer complex demographic histories from DNA sequence data. Most, if not all, of these inference methods exhibit pathological behavior, however. Specifically, they often display runaway behavior in optimization, where the inferred population sizes and epoch durations can degenerate to 0 or diverge to infinity, and show undesirable sensitivity to perturbations in the data. The goal of this paper is to provide theoretical insights into why uch problems arise. To this end, we characterize the geometry of the expected SFS for piecewise-constant demographies and use our results to show that the aforementioned pathological behavior of popular inference methods is intrinsic to the geometry of the expected SFS. We provide explicit descriptions and visualizations for a toy model, and generalize our intuition to arbitrary sample sizes using tools from convex and algebraic geometry. We also develop a universal characterization result which shows that the expected SFS of a sample of size n under an *arbitrary* population history can be recapitulated by a piecewise-constant demography with only κ_n epochs, where κ_n is between $n/2$ and $2n-1$. The set of expected SFS for piecewise-constant demographies with fewer than κ_n epochs is open and non-convex, which causes the above phenomena for inference from data.

KEYWORDS population size; expected sample frequency spectrum; coalescent theory; algebraic methods

Introduction

The sample frequency spectrum (SFS), also known as the site or allele frequency spectrum, is a fundamental statistic in population genomics for summarizing the genetic variation in a sample of DNA sequences. Given a sample of n sequences from a panmictic (i.e., randomly mating) population, the SFS is a vector of length $n-1$ of which the kth entry corresponds to the number of segregating sites each with k mutant (or derived) alleles and $n-k$ ancestral alleles. The SFS provides a concise way to summarize n sequences of arbitrary length into just $n-1$ numbers, and is frequently used in empirical population genetic studies to test for deviations from equilibrium models of evolution. For instance, the SFS has been widely used to infer demographic history where the effective population size has changed over time (Nielsen 2000; Gutenkunst *et al.* 2009; Gravel *et al.* 2011; Keinan and Clark 2012; Excoffier *et al.* 2013; Bhaskar *et al.* 2015), and to test for selective neutrality (Kaplan *et al.* 1989; Achaz 2009). In fact, many commonly used population genetic statistics for testing neutrality, such as Watterson's θ_W (Watterson 1975), Tajima's θ_π (Tajima 1983), and Fu and Li's θ_{FL} (Fu and Li 1993) can be expressed as linear functions of the SFS (Durrett 2008).

In the coalescent framework (Kingman 1982b,c,a), the *unnormalized expected* SFS $\boldsymbol{\xi}_n$ for a random sample of n genomes drawn from a population is obtained by taking the expectation of the SFS over the distribution of sample genealogical histories under a specified population demography. In this work, we will be concerned with well-mixed, panmictic populations with

doi: 10.1534/genetics.XXX.XXXXXX
Manuscript compiled: Monday 30th July, 2018

[1]These authors contributed equally to this work.

[2]Corresponding author: Department of Statistics, 321 Evans Hall #3860, University of California, Berkeley, Berkeley, CA 94720-3860. E-mail: yss@berkeley.edu

time-varying historical population sizes, evolving according to the neutral coalescent process with the infinite-sites model of mutation. The coalescent arises as the continuum limit of a large class of discrete models of random mating, such as the Wright-Fisher, Moran, and Cannings exchangeable family of models (Möhle and Sagitov 2001), by a suitable rescaling of time and taking the population size to infinity. The infinite-sites model postulates that every mutation in the genealogy of a sample occurs at a distinct site, and is commonly employed in population genetic studies for organisms with low population-scaled mutation rates, such as humans. The SFS also appears in the context of statistical modeling as a vector of probabilities. In particular, the *normalized expected* SFS $\widehat{\boldsymbol{\xi}}_n$, defined by normalizing the entries of $\boldsymbol{\xi}_n$ so that they sum to 1, gives the probability that a mutation chosen at random is present in k out of n sequences in the sample. Unless stated otherwise, we use the term expected SFS to refer to the unnormalized quantity $\boldsymbol{\xi}_n$.

The expected SFS is strongly influenced by the demographic history of the population, and extensive theoretical and empirical work has been done to characterize this dependence (Fu 1995; Wakeley and Hey 1997; Polanski *et al.* 2003; Marth *et al.* 2004; Chen 2012; Kamm *et al.* 2017; Jouganous *et al.* 2017). Fu (1995) showed that under the infinite-sites model for a panmictic population with constant size and no selection, the expected SFS is given by $\boldsymbol{\xi}_n = \theta \cdot (1, \frac{1}{2}, \ldots, \frac{1}{n-1})$, where $\theta/2$ denotes the population-scaled mutation rate. When the population size is variable, however, the formula for the expected SFS depends on the entire population size history. In particular, Polanski and Kimmel (2003, Equations 13-15) showed that the expected SFS under a time-varying population size is given by $\boldsymbol{\xi}_n = A_n \mathbf{c}$, with A_n being an $(n-1)$-by-$(n-1)$ invertible matrix that only depends on n (formula presented in **Appendix**), and $\mathbf{c} = (c_2, \ldots, c_n)$, where c_m denotes the expected time to the first coalescence event in a random sample of size m drawn from the population at present. For any time-varying population size function $\eta(t)$, the quantity c_m is given by the following expression:

$$c_m = \int_0^\infty \binom{m}{2} \frac{1}{\eta(t)} \exp\left[-\binom{m}{2} \int_0^t \frac{1}{\eta(x)} \mathrm{d}x\right] \mathrm{d}t. \quad (1)$$

Pathologies of SFS-based inference algorithms

Let us consider a hypothetical scenario. Suppose we would like to learn about the population history of a group of finches on a remote island. Fossil evidence indicates that the island experienced many generations with ample resources leading to a large roughly-constant population size. Then, some catastrophe occurred, rendering the island's resources scarce, leading to a small constant population size until the present. We are given four haplotypes from the population, and we hope to infer the following parameters for a demographic model based on the history described above:

1. How big was the population during the epoch of plenty?
2. How big was the population during the epoch of scarcity?
3. When did the catastrophe occur, marking the breakpoint?

First, we compute the SFS for the four haplotypes we collected. (Our choice of sample size four is for simplicity of this example, but the principles apply for larger samples.) We count singleton (appearing in only one of the haplotypes), doubleton, and tripleton mutations. We do not attempt to track non-segregating sites. Now we have the SFS, a vector of three real numbers.

Next, we ask ourselves: would we expect to obtain this SFS for some particular set of parameters, based on our model? If the answer is yes, then that set of parameters is our best guess. In Figure 1, the green region describes the set of SFS we would expect for various parameters under this model. Blue dots indicate measured SFS. When the blue dots land in the green region, we simply infer the parameters corresponding to that point. The red crosses are the expected SFS computed for those parameters, so they coincide with the blue dots.

What if the answer is no? That is, what if the SFS we measured would not be expected for *any* choice of parameters in our population history model? We have two options to interpret this situation: 1) Statistical noise is making the SFS appear inconsistent with the model. 2) Our model is mis-specified. Let's suppose that noise is the culprit. Then our strategy is to look for the *closest* SFS that would be expected in our model, and infer the parameters associated with that one.

This runs into two problems: First off, the parameters inferred in this way are often nonsensical. In Figure 1, the blue dots outside of the green region are connected by dotted lines to the closest SFS vectors in the green region. Naturally, these mainly lie on the boundary of the green region. The problem is that the boundary points (with one exception that we will discuss later) do not actually correspond to achievable expected SFS vectors! Those points correspond to population size histories where one of the epochs is ∞ or 0.

The second problem: even though there is, in general, a unique closest SFS to a given point outside of the green region, the process of finding the closest point is *highly sensitive* to noise. Specifically, if you change the quantities in the vector by a small amount, the resulting "closest point" may change by a large amount. The reason for this is that the set is *non-convex*, meaning that not all of the straight lines between points in the green region lie inside the green region. As a consequence, some of the blue dots point to the left-hand green region, while others nearby point to the right-hand green region. Sensitivity to noise is a big problem for inference. Any demographic inference method would manifest these pathologies; indeed, the commonly used ∂a∂i (Gutenkunst *et al.* 2009), `fastsimcoal2` (Excoffier *et al.* 2013), and `fastNeutrino` (Bhaskar *et al.* 2015) all encounter these issues.

If we hypothesize that the model may be mis-specified, we need to support this assertion. The question will arise, "How *far away* is our measured SFS from the type of SFS that we would expect under the rejected population model?" Furthermore, we may be asked to offer an alternative hypothesis, i.e. is there another model that actually does allow for an SFS equal to or near the one that we measured? Both of these questions require an understanding of the set of all possible SFS.

Minimal demographic complexity for SFS reconstruction

Let us slightly change our finch example. Suppose we have no *a priori* assumptions regarding the demographic history. Instead, we are only interested in determining whether the SFS is consistent with a null hypothesis of a single panmictic population under neutrality. If the measured SFS is equal to the expected SFS for some demography, we may be asked to produce the *simplest* demography with the expected SFS we want. Work by Myers *et al.* (2008) implies that there are infinitely many population size histories with a given expected SFS, as long as we allow the demographies to be arbitrarily complicated. The paper Bhaskar and Song (2014) by two of this paper's authors demonstrated that when we constrain ourselves to a simpler family of population size histories, we may have a unique function achieving the desired expected SFS.

Now suppose that the SFS does not equal the expected SFS for *any* demography. Again, we would need to quantify how far away it is from being achieved by some demography. This is an intimidating task. How can we be certain to find the SFS corresponding to every demography without leaving any SFS vectors out? After all, the space of possible population size histories is infinite-dimensional! Our hope is to understand the *shape* of the set of all possible SFS vectors so we know that we have covered everything when we reject the null hypothesis.

For the small example of sample size 4, we have demonstrated a sequence of constraints placed on SFS vectors in Figure 2. The vectors of interest have three coordinates corresponding to singleton, doubleton, and tripleton mutations. Note that any vector of probabilities must be non-negative, and must sum to 1. This means we are constrained to the triangle with vertices $(1,0,0)$, $(0,1,0)$, and $(0,0,1)$. We can ignore the third coordinate, since it will always be one minus the others. This triangle is depicted in yellow in Figure 2. One might naively hope that every one of these probability vectors is achievable as the expected SFS of some demography.

A result proved by Sargsyan and Wakeley (2008) is that SFS vectors must be non-increasing–this means we are left with the triangle with vertices $(1/3,1/3,1/3)$, $(1/2,1/2,0)$, and $(1,0,0)$. This is depicted in blue in Figure 2. They further proved that the SFS is convex. This implies that the second coordinate is less than the average of the other two. This further cuts down our possibilities to the triangle with vertices $(1/3,1/3,1/3)$, $(2/3,1/3,0)$, and $(1,0,0)$, depicted in red in Figure 2. If we want SFS vectors for population size histories with two constant pieces, we are further constrained to the green region, which we will describe algebraically later.

We will be able to completely describe the shape of all SFS for sample size 4 using algebraic formulae for the boundary. In fact, we will show that to find all possible SFS for sample size 4, it is sufficient to consider piecewise-constant functions with at most three constant pieces! Furthermore, we will use tools from *convex* and *algebraic geometry* to extend our intuition from this small case study to the SFS for all sample sizes.

Summary of main results

Studying the geometry of the set of expected SFS will address both of the areas discussed above:

1. Explaining the pathologies in SFS-based inference, and
2. describing the full set of SFS for fixed sample size.

In this way, we can help researchers understand why fitting parameters to certain demographic models will lead to runaway behavior. We also enable researchers to reject a null-hypothesis of a single panmictic population under neutrality.

Our main result is Theorem 8, which focuses on piecewise-constant demographies. It shows that for every sample size n, there is a crucial threshold in demographic complexity, which we denote κ_n. If we are fitting to a demographic model with fewer than κ_n constant pieces, then the set of all SFS will be *non-convex* and we must expect pathological behavior as described above. Once we allow for κ_n constant pieces, though, we get the *full* set of SFS for *all* demographies. Proving that this set is convex is left for later work.

Piecewise-Constant Demographies

In this section, we will define two sets: one of them will be the set of expected SFS for piecewise-constant population size histories. As described in the introduction, this is an important set for inference. The other set is the set of expected coalescence vectors; this is not as commonly-used as the SFS, but it helps us build a strong understanding of the SFS. This is because it is related to the SFS by a simple transformation, and yet it is much easier to formulate.

Let Π_k be the set of piecewise-constant population size functions with k pieces. Any population size function in Π_k is described by $2k-1$ positive numbers, representing the k population sizes $(y_1,\ldots,y_k)$ and the $k-1$ time points $(t_1,\ldots,t_{k-1})$ when the population size changes. Let $\Xi_{n,k}$, which we call the (n,k)-SFS manifold[1], denote the set of all expected SFS vectors for a sample of size n that can be generated by population size functions in Π_k. Similarly, let $\mathcal{C}_{n,k}$, called the (n,k)-coalescence manifold, denote the set of all vectors $\mathbf{c}=(c_2,\ldots,c_n)$ giving the expected first coalescence times of samples of size $2,\ldots,n$ for population size functions in Π_k. Let $\widehat{\Xi}_{n,k}$ and $\widehat{\mathcal{C}}_{n,k}$ respectively be equal to the normalization of all points in $\Xi_{n,k}$ and $\mathcal{C}_{n,k}$ by their ℓ_1-norms (i.e., the sums of their coordinates). Note that both manifolds live in $\mathbb{R}^{n-1}$ and their normalized versions live in the $(n-2)$-dimensional simplex Δ^{n-2}; this is the set of nonnegative vectors in $\mathbb{R}^{n-1}$ whose coordinates sum to 1.

Now that we have defined our basic objects of study, we can describe the remainder of the paper: First, we provide a complete geometric picture of the $\Xi_{4,k}$ SFS manifold describing the expected SFS for samples of size $n=4$ under piecewise-constant population size functions with an arbitrary number k of pieces. We make explicit the map between regions of the demographic model space and the corresponding probability vectors, and this will foreshadow some of the difficulties with population size inference in practice. Next, we develop a characterization of the space of expected SFS for arbitrary population size histories. In particular, we show that for any sample size n, there is a finite integer κ_n such that the expected SFS for a sample of n under any population size history can be generated by a piecewise-constant population size function with at most κ_n epochs. Stated another way, we show that the Ξ_{n,κ_n} SFS manifold contains the expected SFS for all possible population size histories, no matter how complicated their functional forms. We establish bounds on κ_n that are linear in n, and along the way prove some interesting results regarding the geometry of the general $\Xi_{n,k}$ SFS manifold.

Before proceeding further, we state a proposition regarding the structure of the map from Π_k to $\mathcal{C}_{n,k}$, which we will call $\chi(\vec{x},\vec{y})$; the vector of $k-1$ transformed breakpoints is denoted by $\vec{x}=(x_1,\ldots,x_{k-1})$ and defined below, while the vector of population sizes in the k epochs is denoted by $\vec{y}=(y_1,\ldots,y_k)$. It turns out that we can formulate the expected coalescence times as polynomial functions of the x and y variables. Two different ways of writing those functions down will give us two perspectives on their shape. All proofs of the results presented in this paper are deferred to **Appendix**.

Proposition 1. *Fix a piecewise-constant population size function in Π_k with epochs $[t_0,t_1)$, $[t_1,t_2),\ldots,[t_{k-1},t_k)$, where $0=t_0<t_1<\cdots<t_{k-1}<t_k=\infty$, and which has constant population size value y_j in the epoch $[t_{j-1},t_j)$ for $j=1,\ldots,k$. Let $x_j=\exp[-(t_j-t_{j-1})/y_j]$ for $j=1,\ldots,k$, where $x_k=0$ (corresponding to time $T=$*

[1] The sets $\Xi_{n,k}$ and $\mathcal{C}_{n,k}$ are not technically manifolds; they would be more accurately described as semialgebraic sets. However, for expository purposes, we use the widely known term "manifold."

∞), and define $x_0 = 1$ (corresponding to time $T = 0$) for convenience. The vectors $(x_1, \ldots, x_{k-1}, y_1, \ldots, y_k)$, where $0 < x_j < 1$ and $y_j > 0$ for all j, (uniquely) identify the population size functions in Π_k, and they satisfy both of the following equations:*

$$\begin{bmatrix} x_0(1-x_1) & \ldots & \left(\prod_{i=0}^{k-1} x_i\right)(1-x_k) \\ \frac{1}{3}x_0^3(1-x_1^3) & \ldots & \frac{1}{3}\left(\prod_{i=0}^{k-1} x_i^3\right)(1-x_k^3) \\ \vdots & \ddots & \vdots \\ \frac{1}{\binom{n}{2}}x_0^{\binom{n}{2}}\left(1-x_1^{\binom{n}{2}}\right) & \ldots & \frac{1}{\binom{n}{2}}\left(\prod_{i=0}^{k-1} x_i^{\binom{n}{2}}\right)\left(1-x_k^{\binom{n}{2}}\right) \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_k \end{bmatrix} = \begin{bmatrix} c_2 \\ c_3 \\ \vdots \\ c_n \end{bmatrix}, \quad (2)$$

$$\begin{bmatrix} 1 & x_1 & \ldots & \prod_{i=1}^{k-1} x_i \\ \frac{1}{3} & \frac{1}{3}x_1^3 & \ldots & \frac{1}{3}\prod_{i=1}^{k-1} x_i^3 \\ \vdots & \vdots & \ddots & \vdots \\ \frac{1}{\binom{n}{2}} & \frac{1}{\binom{n}{2}}x_1^{\binom{n}{2}} & \ldots & \frac{1}{\binom{n}{2}}\prod_{i=1}^{k-1} x_i^{\binom{n}{2}} \end{bmatrix} \begin{bmatrix} y_1 \\ y_2-y_1 \\ \vdots \\ y_k-y_{k-1} \end{bmatrix} = \begin{bmatrix} c_2 \\ c_3 \\ \vdots \\ c_n \end{bmatrix}, \quad (3)$$

where c_m is the expected first coalescence time for a sample of size m, as defined in (1).

These two formulations provide two different ways of looking at the coalescence manifold $\mathcal{C}_{n,k}$:

1. In (2), the left-hand matrix, called $M_1(n,k)$, has each column of the same form with two parameters; this indicates they all live in a 2-dimensional surface. Imagine, for example, the surface of the earth. There are two degrees of freedom: north-south and east-west. Here, too, specifying the value of each column, regardless of the value of n, is dependent on two numbers. Explicitly, each column is given by $f_n(a,b) = \left(a(1-b), \ldots, a^{\binom{n}{2}}(1-b^{\binom{n}{2}})/\binom{n}{2}\right)$ for some inputs a and b.

   Additionally, the vector $(y_1, \ldots, y_k)$ has all positive entries. That means that, when we combine columns from our surface, they will not cancel in unexpected ways due to negative coefficients. The set of positive combinations of a set of points is called a cone, and it is very nicely behaved geometrically. This means that the vector $\mathbf{c} = (c_2, \ldots, c_n)$ is contained in the cone over the surface described by the columns of M_1.

2. In (3), the left-hand matrix, call it $M_2(n,k)$ has each column of the same form with one parameter; this indicates they all live on a curve. Like a train on a track, this has one degree of freedom, only forward-backward. Explicitly, each column is given by $g_n(a) = \left(a, \ldots, a^{\binom{n}{2}}/\binom{n}{2}\right)$ for some input a.

   The vector $(y_1, y_2-y_1, \ldots, y_k-y_{k-1})$ on the left hand side has entries with possibly negative coordinates. So the vector $\mathbf{c} = (c_2, \ldots, c_n)$ is contained in the linear span of the curve described by the columns of M_2. Unfortunately, a linear span is not quite as nicely behaved as a cone. Still, this formulation gains the simplicity of having *one* degree of freedom instead of two.

Proposition 1 gives us the algebraic mappings that will serve as our objects of interest. Since the SFS manifold is simply a linear transformation of the coalescence manifold, we will use these maps as our entry into understanding the SFS manifold.

The $\Xi_{4,k}$ SFS Manifold: A Toy Model

The first in-depth study will involve the set of all possible expected SFS for a sample of size 4. We choose $n = 4$ for a number of reasons: First, the cases of sample size 2 and 3 are not interesting. When we only have two haplotypes, there is only one entry in the SFS vector, i.e. singletons. The resulting set of possible expected SFS is just the set of all positive numbers. When we have three haplotypes, it's only slightly better. Because there must be fewer doubletons than singletons, the possible expected SFS is somewhere in the wedge between 0° and 45° from the origin; this turns out to be the only constraint.

Second, when $n = 4$, the SFS manifold lives in $\mathbb{R}^3$, which can be nicely visualized, and the normalized SFS manifold lives in the triangle with vertices $(1,0,0)$, $(0,1,0)$, and $(0,0,1)$. Finally, as observed in Proposition 1, the most interesting phenomena in SFS manifolds of any dimension are fundamentally phenomena of curves and surfaces. These are already captured in the $n = 4$ case.

For the sake of completeness, we begin by formally describing the coalescence manifolds $\mathcal{C}_{n,k}$ for the trivial cases of $n = 2$ and $n = 3$.

Proposition 2. *We list some basic results on the coalescence manifolds $\mathcal{C}_{n,k}$, with sample size n and k population epochs, for small values of (n,k):*

1. *The manifold* $\mathcal{C}_{n,1} = \left\{\lambda \cdot \left(1, \frac{1}{3}, \ldots, \frac{1}{\binom{n}{2}}\right) : \lambda > 0\right\}$*, for all n.*
2. *The manifold* $\mathcal{C}_{2,k} = \mathcal{C}_{2,1} = \{a : a > 0\}$*, for all $k \geq 1$.*
3. *The manifold* $\mathcal{C}_{3,k} = \mathcal{C}_{3,2} = \{(a,b) : a > 0 \text{ and } 0 < b < a\}$*, for all $k \geq 2$.*

Note that from (2) and (3) for $\chi(\vec{x}, \vec{y})$, it follows that $\chi(\vec{x}, a\vec{y}) = a\chi(\vec{x}, \vec{y})$ for $a > 0$. In words, rescaling the population sizes in each epoch by a constant a also rescales the first coalescence times by a. This implies that every point in the coalescence manifold $\mathcal{C}_{n,k}$ generates a full ray contained in the $\mathcal{C}_{n,k}$ coalescence manifold. Another consequence is that the normalized coalescence manifold $\widehat{\mathcal{C}}_{n,k}$ is precisely the intersection of the coalescence manifold $\mathcal{C}_{n,k}$ with the simplex Δ^{n-2}.

With that justification, we begin to consider the normalized coalescence manifold $\widehat{\mathcal{C}}_{4,k}$ living in the simplex. As stated in Proposition 2, $\mathcal{C}_{4,1}$ is a ray, which implies that $\widehat{\mathcal{C}}_{4,1}$ is a single point. We now characterize the set $\widehat{\mathcal{C}}_{4,2}$. Again, this is the set of possible SFS for two-epoch piecewise-constant population size histories considered as a subset of all vectors summing to one.

Proposition 3. *The manifold $\widehat{\mathcal{C}}_{4,2}$, describing normalized expected times to first coalescence for sample size 4 and two population epochs, is a two-dimensional subset of the 2-simplex which can be described as the union of the point $\widehat{\mathcal{C}}_{4,1}$ with the interiors of the convex hulls of two curves γ_1 and γ_2. The curves are parametrized as follows:*

$$\gamma_1 = \left\{\left(\frac{6}{6+2t^2+t^5}, \frac{2t^2}{6+2t^2+t^5}, \frac{t^5}{6+2t^2+t^5}\right) : 0 < t < 1\right\},$$

$$\text{and} \quad \gamma_2 = \left\{\left(\frac{6}{6+2[2]_t+[5]_t}, \frac{2[2]_t}{6+2[2]_t+[5]_t}, \frac{[5]_t}{6+2[2]_t+[5]_t}\right) : 0 < t < 1\right\},$$

where $[n]_t$ denotes $1 + \cdots + t^n$.

This set has some highly unpleasant geometry. First of all, the set is non-convex; topologically, it is also neither closed nor open, because most of the boundary is excluded with the exception of the point $(2/3, 2/9, 1/9)$. The set is visualized in Figure 3A.

In order to precisely illustrate the geometry of $\chi(\vec{x},\vec{y})$, we will consider how contours in the domain map to contours in the image. Specifically, we plot the images of lines with fixed values of x_1, respectively fixed values of (y_1, y_2), to $\mathcal{C}_{4,2}$ in the 2-simplex. The resulting contours are pictured in Figure 4.

Finally, we consider how the map χ acts on the boundaries of the domain. To aid visualization, we limit the inputs to x_1 and y_1/y_2, since all rescalings of y_1 and y_2 by the same positive constant while keeping x_1 fixed map to the same normalized coalescence vector. The resulting map is illustrated in Figure 5.

We note that the map fails to be one-to-one within the domain only when $y_1/y_2 = 1$; this is also in the pre-image of the point $(\frac{2}{3}, \frac{2}{9}, \frac{1}{9}) \in \widehat{\mathcal{C}}_{4,2}$. The inverse function theorem implies that on the complement of $y_1/y_2 = 1$, the map is a homeomorphism (a map that preserves topological features like number of components). This is consistent with our observation that the two rectangles in Figure 5A correspond to the two envelopes in Figure 5C. Now, we consider demographies with more than two epochs. This proposition implies that any expected SFS for sample size 4 coming from a single panmictic population under neutrality, regardless of the true population size history, is *equal* to the expected SFS for some piecewise-constant history with only three pieces. It also shows that all of these SFS vectors live inside of the convex hull of one curve.

Proposition 4. *For all values $k \geq 3$, the manifold $\widehat{\mathcal{C}}_{4,k} = \widehat{\mathcal{C}}_{4,3}$, and $\widehat{\mathcal{C}}_{4,3}$ is the interior of the convex hull of the following curve:*

$$\gamma_3 = \left\{ \left(\frac{1}{1+t^2+t^5}, \frac{t^2}{1+t^2+t^5}, \frac{t^5}{1+t^2+t^5} \right) : 0 < t < 1 \right\}.$$

As we can see from Proposition 4, $\widehat{\mathcal{C}}_{4,3}$ is open and convex; however, we lose one useful property of the normalized map $\widehat{\chi} : \mathbb{R}^3 \to \widehat{\mathcal{C}}_{4,2}$. Specifically, let $\widehat{\chi}' : \mathbb{R}^2 \to \widehat{\mathcal{C}}_{4,2}$ be given by $\widehat{\chi}'(x_1, y_1) = \widehat{\chi}(x_1, y_1, 1)$, noting that $\widehat{\chi}(x_1, \lambda y_1, \lambda y_2) = \widehat{\chi}(x_1, y_1, y_2)$ for $\lambda > 0$. Under this definition $\widehat{\chi}'$ is generically one-to-one (i.e., one-to-one away from a set of measure zero). Meanwhile, the analogous construction $\widehat{\chi}' : \mathbb{R}^4 \to \widehat{\mathcal{C}}_{4,3}$ mapping the three-epoch demography with breakpoints (x_1, x_2) and population sizes $(y_1, y_2, 1)$ to the corresponding normalized coalescence vector has two-dimensional pre-images, generically. For this reason, contour images do not lend themselves to easy description.

However, as a heuristic, we can choose a distinguished member of this pre-image with nice properties. In the orange region adjacent to β_3 depicted in Figure 6, every pre-image contains a limit demography with first and third epochs set to zero, and second epoch set to one. This can be thought of as a demography with a population boom in the second epoch. In the blue region adjacent to the line segment from $(1/3, 1/3, 1/3)$ to $(1, 0, 0)$, every pre-image contains a limit demography with second epoch set to zero. This corresponds to a demography with a population bottleneck in the second epoch. Because the set of demographies mapping to each point is two-dimensional, this does not describe all demographies characterized by a chosen SFS, but it does give us intuition for the types of demographies to expect.

We can also describe the image of the map $\widehat{\chi}' : \mathbb{R}^4 \to \widehat{\mathcal{C}}_{4,3}$ on the boundaries of our domain. The easiest way to visualize the map is first to understand how the time variables affect the value of the columns of $M_1(4,3)$ and to view the y variables as specifying points in the convex hull of those 3 columns. The boundaries of the square $(x_1, x_2) \in [0,1] \times [0,1]$ map the columns (after rescaling to the simplex) as follows:

$$x_1 = 0 \quad \mapsto \quad \left[\begin{array}{c|c|c} 6/9 & 1 & 1 \\ 2/9 & 0 & 0 \\ 1/9 & 0 & 0 \end{array} \right],$$

$$x_1 = 1 \quad \mapsto \quad \left[\begin{array}{c|c|c} 1/3 & | & | \\ 1/3 & \gamma_2(x_2) & \gamma_1(x_2) \\ 1/3 & | & | \end{array} \right],$$

$$x_2 = 0 \quad \mapsto \quad \left[\begin{array}{c|c|c} | & | & 1 \\ \gamma_2(x_1) & \gamma_1(x_1) & 0 \\ | & | & 0 \end{array} \right],$$

$$x_2 = 1 \quad \mapsto \quad \left[\begin{array}{c|c|c} | & | & | \\ \gamma_2(x_1) & \gamma_3(x_1) & \gamma_1(x_1) \\ | & | & | \end{array} \right].$$

The case of $x_2 = 1$ is the most interesting: when we fix $y_1 = y_3 = 0$ and $y_2 = 1$, we obtain the boundary curve $\gamma_3(t)$. Note that $x_2 = 1$ corresponds to a second epoch of length 0. The intuition is that very short population booms at the second epoch lead to coalescence vectors close to γ_3. The maps encoded by a general column of $M_1(4,k)$ correspond to the interior of the orange region in Figure 7A. Adding in convex combinations of points gives the lined region, which is the remainder of $\mathcal{C}_{4,3}$; this is discussed more rigorously in **Appendix**. When the number of epochs k steps higher, all columns of $M_1(4,k)$ still map to the same region of the simplex, so $\mathcal{C}_{4,k}$ will still be contained in this convex hull. The region $\mathcal{C}_{4,3}$ is depicted in Figure 7A.

As mentioned earlier, the SFS manifold $\Xi_{n,k}$ is merely a linear transformation of $\mathcal{C}_{n,k}$; however, since it is of interest in its own right, we include the formulae for $\Xi_{4,k}$ analogous to those derived in this section.

Proposition 5. *The following hold for the normalized $(4,k)$-SFS manifold:*

$$\widehat{\Xi}_{4,1} = \left(\frac{6}{11}, \frac{3}{11}, \frac{2}{11} \right).$$

$\widehat{\Xi}_{4,2}$ is the union of $\widehat{\Xi}_{4,1}$ with the convex hulls of two curves:

$$\beta_1 = \left\{ \left(\frac{18+10t^2+2t^5}{54+t^5}, \frac{18-3t^5}{54+t^5}, \frac{18-10t^2+2t^5}{54+t^5} \right) : 0 < t < 1 \right\},$$

$$\beta_2 = \left\{ \left(\frac{18+10[2]_t+2[5]_t}{54+[5]_t}, \frac{18-3[5]_t}{54+[5]_t}, \frac{18-10[2]_t+2[5]_t}{54+[5]_t} \right) : 0 < t < 1 \right\}.$$

Here, also, $[n]_t$ denotes $1 + t + \cdots + t^n$. Finally, $\widehat{\Xi}_{4,k} = \widehat{\Xi}_{4,3}$ for all k, and $\widehat{\Xi}_{4,3}$ is the convex hull of β_3, where

$$\beta_3 = \left\{ \left(\frac{3+5t^2+2t^5}{9+t^5}, \frac{3-3t^5}{9+t^5}, \frac{3-5t^2+2t^5}{9+t^5} \right) : 0 < t < 1 \right\}.$$

Visualizations of $\Xi_{4,2}$ and $\Xi_{4,3}$ may be found in Figure 3B and Figure 7B.

The $\Xi_{n,k}$ SFS Manifold: General Properties

In this section, we examine the constant κ_n, defined earlier as the smallest index for which $\mathcal{C}_{n,k} \subseteq \mathcal{C}_{n,\kappa_n}$ for all k. The tools for the proofs in this section come from algebraic geometry (for the derivation of the lower bound) and convex geometry (for the upper bound).

The gist of the algebraic geometry argument is that, under the $M_2(n,k)$ formulation, the manifold $\mathcal{C}_{n,k}$ can be seen to be part of another manifold built by a sequence of well-understood algebraic constructions. Details of this perspective are reserved for the Proofs section.

Two concrete consequences follow from this observation:

1. the ability to compute all equations satisfied by $\mathcal{C}_{n,k}$ using computer algebra, and
2. a formula for the dimension of the coalescence and SFS manifolds.

While the former is harder to explain without more setup, the latter can be simply stated: the dimension of the normalized coalescence manifold $\widehat{\mathcal{C}}_{n,k}$ is 0 when we have the constant demography ($k=1$). If we allow k constant pieces, the manifold has dimension $2k-2$ unless $2k-2$ is greater than $n-2$, the dimension of the simplex Δ^{n-2}. In that case, it has dimension $n-2$.

Proposition 6. *The dimension of $\widehat{\mathcal{C}}_{n,k}$ is given by:*

$$\dim \widehat{\mathcal{C}}_{n,k} = \begin{cases} 0, & k=1, \\ \min(2k-2, n-2), & \textit{else.} \end{cases}$$

In particular, $\mathcal{C}_{n,k}$ is a proper subset of $\mathcal{C}_{n,k+1}$ for $k < \lceil \frac{1}{2} n \rceil$.

While Proposition 6 is useful for analyzing individual coalescence manifolds, it also leads to the observation that $\kappa_n \geq \lceil \frac{1}{2} n \rceil$, since the inclusions are proper until that index. It is worth remarking that a slightly weaker lower bound of $\kappa_n \geq \lfloor \frac{1}{2} n \rfloor$ follows immediately from the identifiability result of Bhaskar and Song (2014, Corollary 7), which states that for a piecewise-constant population size function with k pieces, the expected SFS of a sample of size $n \geq 2k$ suffices to uniquely identify the function.

We will illustrate how these algebraic ideas can be applied in the next case we have not seen, namely sample size $n=5$.

Example 7. Note that $\widehat{\mathcal{C}}_{5,1} = \left(\frac{30}{48}, \frac{10}{48}, \frac{5}{48}, \frac{3}{48} \right)$, by Proposition 2. We will use the new ideas above to describe $\widehat{\mathcal{C}}_{5,k}$ for higher values of k.

Since the normalized coalescence manifold has dimension $\min(2k-2, n-2)$, we know that $\widehat{\mathcal{C}}_{5,2}$ has dimension 2 inside of the 3-simplex; therefore, we anticipate that it will satisfy *one* equation, matching its codimension. The degree of the algebraic variety implies that this polynomial should have degree 8. Indeed, when we compute this equation using computer algebra software `Macaulay2` (Grayson and Stillman 2002), we obtain a huge degree-8 polynomial with 105 terms, whose largest integer coefficient is 5,598,720. Finally, $\widehat{\mathcal{C}}_{5,3}$ is full-dimensional in the 3-simplex, so it will satisfy no algebraic equations relative to the simplex. It would be defined instead by the inequalities determining its boundary.

The convex geometry argument is more elementary. As we noted, the M_1 formulation is contained in the convex hull over the surface described by a general column of M_1. Because the columns are related, our selection of points in the surface is not unrestricted. For this reason, it is not obviously *equal* to the convex hull. However, once we fix some collection of values $x_1, \ldots, x_k$ to be input in the formula for $\mathcal{C}_{n,k}$, we can use convex geometry for the resulting polytope. In particular, we use Caratheodory's Theorem (Carathéodory (1907) or Barvinok (2002, Theorem 2.3)), which states that for X a subset of $\mathbb{R}^n$, every $x \in \text{cone}(X)$ can be represented as a positive combination of vectors $x_1, \ldots, x_m \in X$ for some $m \leq n$.

The argument, roughly, allows us to construct any point in that convex hull, with as few as $n+1$ points. This allows us to place the point in $\mathcal{C}_{n,j}$ for $j \leq 2n-1$. Since no new SFS are generated by using more than $2n-1$ epochs, we learn that κ_n is bounded above by $2n-1$.

Combining the two bounds obtained in this section, we have the main theorem described in the Introduction.

Theorem 8. *For any integer $n \geq 2$, there exists a positive integer κ_n such that $\Xi_{n,k} \subseteq \Xi_{n,\kappa_n}$ for all $k \geq 1$. Furthermore, κ_n satisfies*

$$\lceil n/2 \rceil \leq \kappa_n \leq 2n-1.$$

Additionally, $\Xi_{n,k}$ is nonconvex for all values of $2 \leq k < \kappa_n$.

This allows us to express the SFS from any piecewise-constant demography as coming from a demography with relatively few epochs. Because the SFS is an integral over the demography, the SFS from a general measurable demography can be uniformly approximated by a piecewise-constant demography with sufficiently many epochs. Our results imply that it can be precisely obtained by a demography with at most $2n-1$ epochs.

Discussion

In this work, we characterized the manifold of expected SFS $\Xi_{n,k}$ generated by piecewise-constant population histories with k epochs, while giving a complete geometric description of this manifold for the sample size $n=4$ and $k=2$ epochs. This special case is already rich enough to shed light on the issues that practitioners can face when inferring population demographies from SFS data using popular software programs. While we demonstrated these issues in Figure 1 using the `fastNeutrino` program (Bhaskar *et al.* 2015), the issues we point out are *inherent* to the geometry of the SFS manifold and not specific to any particular demographic inference software. Our simulations showed that the demographic inference problem from SFS data can be fraught with interpretability issues, due to the sensitivity of the inferred demographies to small changes in the observed SFS data. These results can also be viewed as complementary to recent pessimistic minimax bounds on the number of segregating sites required to reliably infer ancient population size histories (Terhorst and Song 2015; Baharian and Gravel 2018).

Our investigation of piecewise-constant population histories also let us show a general result that the expected SFS for a sample of size n under *any population history* can also be generated by a piecewise-constant population history with at most $2n-1$ epochs. This result could have potential applications for developing non-parametric statistical tests of neutrality. Most existing tests of neutrality using classical population genetic statistics such as Tajima's D (Tajima 1989) implicitly test the null hypothesis of selective neutrality *and* a constant effective population size (Stajich and Hahn 2004). We have characterized the expected SFS of samples of size n under arbitrary population histories in terms of the expected SFS under piecewise-constant population histories with at most κ_n epochs. As a result, the KL divergence of an observed SFS $\boldsymbol{\xi}_n^{\text{obs}}$ to the expected SFS $\boldsymbol{\xi}_n(\eta^*)$ under the best-fitting piecewise constant population history $\eta^* \in \Pi_{\kappa_n}$ with at most $\kappa_n \leq 2n-1$ epochs is also equal (up to a constant shift) to the negative log-likelihood of the observed SFS $\boldsymbol{\xi}_n^{\text{obs}}$ under the best fitting population size history *without any constraints on its form*. (This assumes the commonly-used Poisson Random Field model where sites being analyzed are unlinked.) One can then use the KL divergence inferred by existing parametric demographic inference programs to create rejection regions for the null hypothesis of selective neutrality without having to make any parametric assumption on the underlying demography. Such an approach would also obviate the need for interpreting the inferred demography itself, since the space of piecewise-constant population histories is only being used to compute the best possible log-likelihood under any single population demographic model. This approach could serve as an alternative to recent works which first estimate a parametric demography using genome-wide sites, and then perform a hypothesis test in each genomic region using simulated distributions of SFS statistics like Tajima's D under the inferred demography (Rafajlović *et al.* 2014). We leave the exploration of such tests for future work.

Data Availability

The authors affirm that all data necessary for confirming the conclusions of the article are present within the article, figures, and tables.

Acknowledgments

We thank Simon Gravel, Jeremy Berg, Laura Hayward, Yuval Simons, and the referees for their careful reading of our manuscript and for providing us with helpful comments. We also thank the Simons Institute for the Theory of Computing, where some of this work was carried out while the authors were participating in the "Evolutionary Biology and the Theory of Computing" program. This research is supported in part by a Math+X Research Grant, an NSF grant DMS-1149312 (CAREER), an NIH grant R01-GM109454, and a Packard Fellowship for Science and Engineering. YSS is a Chan Zuckerberg Biohub investigator.

Literature Cited

Achaz, G., 2009 Frequency spectrum neutrality tests: one for all and all for one. Genetics **183**: 249–258.

Baharian, S. and S. Gravel, 2018 On the decidability of population size histories from finite allele frequency spectra. Theoretical population biology .

Barvinok, A., 2002 *A course in convexity*, volume 54. American Mathematical Society Providence.

Bhaskar, A. and Y. S. Song, 2014 Descartes' rule of signs and the identifiability of population demographic models from genomic variation data. Annals of Statistics **42**: 2469–2493.

Bhaskar, A., Y. R. Wang, and Y. S. Song, 2015 Efficient inference of population size histories and locus-specific mutation rates from large-sample genomic variation data. Genome Research **25**: 268–279.

Carathéodory, C., 1907 Über den variabilitätsbereich der koeffizienten von potenzreihen, die gegebene werte nicht annehmen. Mathematische Annalen **64**: 95–115.

Chen, H., 2012 The joint allele frequency spectrum of multiple populations: A coalescent theory approach. Theoretical Population Biology **81**: 179–195.

Durrett, R., 2008 *Probability models for DNA sequence evolution*. Springer Science & Business Media.

Excoffier, L., I. Dupanloup, E. Huerta-Sánchez, V. C. Sousa, and M. Foll, 2013 Robust demographic inference from genomic and SNP data. PLoS Genetics **9**: e1003905.

Fu, Y.-X., 1995 Statistical properties of segregating sites. Theoretical Population Biology **48**: 172–197.

Fu, Y.-X. and W.-H. Li, 1993 Statistical tests of neutrality of mutations. Genetics **133**: 693–709.

Gravel, S., B. M. Henn, R. N. Gutenkunst, A. R. Indap, G. T. Marth, *et al.*, 2011 Demographic history and rare allele sharing among human populations. Proceedings of the National Academy of Sciences **108**: 11983–11988.

Grayson, D. R. and M. E. Stillman, 2002 Macaulay 2, a software system for research in algebraic geometry.

Gutenkunst, R. N., R. D. Hernandez, S. H. Williamson, and C. D. Bustamante, 2009 Inferring the joint demographic history of multiple populations from multidimensional snp frequency data. PLoS Genetics **5**: e1000695.

Harris, J., 2013 *Algebraic geometry: a first course*, volume 133. Springer Science & Business Media.

Jouganous, J., W. Long, A. P. Ragsdale, and S. Gravel, 2017 Inferring the joint demographic history of multiple populations: Beyond the diffusion approximation. Genetics **206**: 1549–1567.

Kamm, J. A., J. Terhorst, and Y. S. Song, 2017 Efficient computation of the joint sample frequency spectra for multiple populations. Journal of Computational and Graphical Statistics **26**: 182–194.

Kaplan, N. L., R. Hudson, and C. Langley, 1989 The "hitchhiking effect" revisited. Genetics **123**: 887–899.

Keinan, A. and A. G. Clark, 2012 Recent explosive human population growth has resulted in an excess of rare genetic variants. Science **336**: 740–743.

Kelleher, J., A. M. Etheridge, and G. McVean, 2016 Efficient coalescent simulation and genealogical analysis for large sample sizes. PLoS Computational Biology **12**: e1004842.

Kingman, J. F. C., 1982a The coalescent. Stochastic Processes and Their Applications **13**: 235–248.

Kingman, J. F. C., 1982b Exchangeability and the evolution of large populations. In *Exchangeability in Probability and Statistics*, edited by G. Koch and F. Spizzichino, pp. 97–112, North-Holland Publishing Company.

Kingman, J. F. C., 1982c On the genealogy of large populations. Journal of Applied Probability **19**: 27–43.

Marth, G. T., E. Czabarka, J. Murvai, and S. T. Sherry, 2004 The allele frequency spectrum in genome-wide human variation data reveals signals of differential demographic history in three large world populations. Genetics **166**: 351–372.

Möhle, M. and S. Sagitov, 2001 A classification of coalescent processes for haploid exchangeable population models. Annals of Probability **29**: 1547–1562.

Myers, S., C. Fefferman, and N. Patterson, 2008 Can one learn history from the allelic spectrum? Theoretical Population Biology **73**: 342–348.

Nielsen, R., 2000 Estimation of population parameters and recombination rates from single nucleotide polymorphisms. Ge-

netics **154**: 931–942.
Polanski, A., A. Bobrowski, and M. Kimmel, 2003 A note on distributions of times to coalescence, under time-dependent population size. Theoretical Population Biology **63**: 33–40.
Polanski, A. and M. Kimmel, 2003 New explicit expressions for relative frequencies of single-nucleotide polymorphisms with application to statistical inference on population growth. Genetics **165**: 427–436.
Rafajlović, M., A. Klassmann, A. Eriksson, T. Wiehe, and B. Mehlig, 2014 Demography-adjusted tests of neutrality based on genome-wide SNP data. Theoretical Population Biology **95**: 1–12.
Sargsyan, O. and J. Wakeley, 2008 A coalescent process with simultaneous multiple mergers for approximating the gene genealogies of many marine organisms. Theoretical population biology **74**: 104–114.
Stajich, J. E. and M. W. Hahn, 2004 Disentangling the effects of demography and selection in human history. Molecular Biology and Evolution **22**: 63–73.
Tajima, F., 1983 Evolutionary relationship of DNA sequences in finite populations. Genetics **105**: 437–460.
Tajima, F., 1989 Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics **123**: 585–595.
Terhorst, J. and Y. S. Song, 2015 Fundamental limits on the accuracy of demographic inference based on the sample frequency spectrum. Proceedings of the National Academy of Sciences **112**: 7677–7682.
Wakeley, J. and J. Hey, 1997 Estimating ancestral population parameters. Genetics **145**: 847–855.
Watterson, G., 1975 On the number of segregating sites in genetical models without recombination. Theoretical Population Biology **7**: 256–276.

Appendix

Formula for A_n

Recall that the SFS can be related to times-to-first-coalescence by the formula $\boldsymbol{\xi}_n = A_n \mathbf{c}$. The formula for A_n is given recursively in (Polanski and Kimmel 2003, Equations 13-15) by the following formulae (with variable names changed for clarity):

$$\begin{aligned}(A_n)_{b,2} &= \frac{6}{n+1} \\ (A_n)_{b,3} &= \frac{30(n-2b)}{(n+1)(n+2)} \\ (A_n)_{b,j+2} &= -\frac{(1+j)(3+2j)(n-j)}{j(2j-1)(n+j+1)}(A_n)_{b,j} \\ &\quad + \frac{(3+2j)(n-2b)}{j(n+j+1)}(A_n)_{b,j+1}\end{aligned}$$

Proof of Proposition 1

First, we reduce the integral expression for c_m to a finite sum; then we make appropriate manipulations until we arrive at the desired expressions.

Coalescence in the Wright-Fisher model is an inhomogeneous Poisson process with parameter $\binom{m}{2}/\eta(t)$. Therefore, the probability density of first coalescence at time T is:

$$\begin{aligned}&\mathbb{P}(\text{No Coalescence in } [0,T))\mathbb{P}(\text{Coalescence at time } T) \\ &\qquad = \exp\left[-\int_0^T \frac{\binom{m}{2}}{\eta(t)}\mathrm{d}t\right]\frac{\binom{m}{2}}{\eta(T)}\mathrm{d}t.\end{aligned}$$

Let $R_\eta(t) = \int_0^T \frac{1}{\eta(t)}\mathrm{d}t$. To compute the expected time to first coalescence, we have the integral:

$$\begin{aligned}c_m &= \int_0^\infty t \cdot \frac{\binom{m}{2}}{\eta(t)} \exp\left[-\binom{m}{2}R_\eta(t)\right]\mathrm{d}t \\ &= \int_0^\infty \exp\left[-\binom{m}{2}R_\eta(t)\right]\mathrm{d}t \quad \text{(Integration by Parts)}\end{aligned}$$

Substituting variables, $\tau = R_\eta(t)$, note that $\mathrm{d}t = \eta(R^{-1}(\tau))\mathrm{d}\tau$. Therefore, the integral becomes:

$$c_m = \int_0^\infty \tilde{\eta}(\tau)\exp\left[-\binom{m}{2}\tau\right]\mathrm{d}\tau,$$

where $\tilde{\eta}(\tau) = \eta(R^{-1}(\tau))$.

The population size $\eta(t)$ is a piecewise constant function, whose value $\eta(t) = \eta_j$ if $t_{j-1} \le t < t_j$. As specified in the Proposition, $t_0 = 0$, $t_k = \infty$, and $(y_1, \ldots, y_k)$ is the vector of population sizes. Observe that $\tilde{\eta}(\tau)$ is also piecewise constant. In particular,

$$\tilde{\eta}(\tau) = \begin{cases} y_1, & 0 \le \tau < \frac{t_1}{y_1}, \\ y_2, & \frac{t_1}{y_1} \le \tau < \frac{t_1}{y_1} + \frac{t_2 - t_1}{y_2}, \\ \vdots & \vdots \end{cases}$$

Let $s_j = t_j - t_{j-1}$ for brevity. The resulting formula is:

$$\tilde{\eta}(\tau) = y_j, \qquad \text{for } \sum_{k=1}^{j-1}\frac{s_k}{y_k} \le \tau < \sum_{k=1}^{j}\frac{s_k}{y_k}.$$

We turn the integral into a sum of integrals on the constant epochs:

$$\begin{aligned}c_m &= \int_0^\infty \tilde{y}(\tau)\exp\left[-\binom{m}{2}\tau\right]\mathrm{d}\tau \\ &= \sum_{j=1}^{k}\int_{\sum^{j-1} s_l/y_l}^{\sum^{j} s_l/y_l} y_j \exp\left[-\binom{m}{2}\tau\right]\mathrm{d}\tau \\ &= \sum_{j=1}^{k} y_j \left[\frac{-1}{\binom{m}{2}}\exp\left[-\binom{m}{2}\tau\right]\right]_{\tau=\sum^{j-1} s_l/y_l}^{\tau=\sum^{j} s_l/y_l} \\ &= \frac{1}{\binom{m}{2}}\left\{\sum_{j=1}^{k} y_j \left(\prod_{l=1}^{j-1}\exp\left[-\binom{m}{2}s_l/y_l\right]\right)\right. \\ &\qquad \left.\left(1 - \exp\left[-\binom{m}{2}s_j/y_j\right]\right)\right\}.\end{aligned}$$

We now make the substitution $x_j = \exp\left[-s_j/y_j\right]$. Note that the old restriction $t_{j+1} > t_j > 0$ becomes the new constraint $0 < x_j < 1$. Our formula for the c_m is now:

$$c_m = \frac{1}{\binom{m}{2}}\left[\sum_{j=1}^{k} y_j \left(\prod_{l=1}^{j-1} x_l^{\binom{m}{2}}\right)\left(1 - x_j^{\binom{m}{2}}\right)\right].$$

Noting the linear form of this expression, we factor as a matrix multiplication:

$$\begin{bmatrix} 1 & & & \\ & \frac{1}{3} & & \\ & & \ddots & \\ & & & \frac{1}{\binom{n}{2}} \end{bmatrix} \times \begin{bmatrix} 1 & x_1 & \dots & \prod_{i=1}^{k-1} x_i \\ 1 & x_1^3 & \dots & \prod_{i=1}^{k-1} x_i^3 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_1^{\binom{n}{2}} & \dots & \prod_{i=1}^{k-1} x_i^{\binom{n}{2}} \end{bmatrix}$$

$$\times \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 \\ -1 & 1 & 0 & \ddots & 0 \\ 0 & -1 & 1 & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & -1 & 1 \end{bmatrix} \times \begin{bmatrix} y_1 \\ \vdots \\ y_k \end{bmatrix} = \begin{bmatrix} c_2 \\ \vdots \\ c_n \end{bmatrix}.$$

Combining the first three matrices yields (2); combining the first two and last two separately yields (3). □

Proof of Proposition 2

We justify each equation in turn:

1. As mentioned in the introduction, this is a classical result in population genetics, and can be derived directly from (3).

2. The inclusion $\mathcal{C}_{2,1} \subset \mathcal{C}_{2,k}$ is immediate, so we need only show that any $a \in \mathcal{C}_{2,k}$ satisfies $a > 0$. Using (2), a is written as a sum of products of strictly positive numbers; so $\mathcal{C}_{2,k} \subset \mathcal{C}_{2,1}$.

3. First, we show that $\mathcal{C}_{3,2}$ is the interior of the open cone spanned by $(1,0)$ and $(1,1)$. Fix $y_1 = a/(1-x_1)$ (for a positive) and consider $\chi(x_1, a/(1-x_1), y_2)$:

$$\chi\left(x_1, \frac{a}{1-x_1}, y_2\right) = \begin{bmatrix} a + x_1 y_2 \\ \frac{1}{3}a(1+x_1+x_1^2) + \frac{1}{3}x_1^3 y_2 \end{bmatrix}$$

$$= a\begin{bmatrix} 1 \\ \frac{1}{3}(1+x_1+x_1^2) \end{bmatrix} + x_1 y_2 \begin{bmatrix} 1 \\ \frac{1}{3}x_1^2 \end{bmatrix}.$$

When $x_1 \to 0$, the second vector approaches $(1,0)$; when $x_1 \to 1$, the first vector approaches $(1,1)$. The vectors are in the interior of that cone for all other permissible values of x_1 and y_2. To show that $\mathcal{C}_{3,k} = \mathcal{C}_{3,2}$, note that for larger values of k, the same cone of vectors are produced. In particular, $\chi(x_1, \dots, x_{k-1}, y_1, \dots, y_k)$ yields

$$\sum_{j=1}^{k-1} \left\{ y_j \left(\prod_{i=1}^{j-1} x_i \right) (1-x_j) \begin{bmatrix} 1 \\ \frac{1}{3}(\prod_{i=1}^{j-1} x_i^2)(1+x_j+x_j^2) \end{bmatrix} \right\}$$

$$+ y_k \left(\prod_{i=1}^{k-1} x_i \right) \begin{bmatrix} 1 \\ \frac{1}{3}(\prod_{i=1}^{k-1} x_i^2) \end{bmatrix}.$$

Clearly, the second coordinate of all vectors is bounded between 0 and 1.

Proof of Proposition 3

First we observe that γ_1 and γ_2 are normalizations of the curves defined by parameterizations $(t, \frac{1}{3}t^3, \frac{1}{6}t^6)$ and $(1-t, \frac{1}{3}(1-t^3), \frac{1}{6}(1-t^6))$ where t is constrained to the open interval $(0,1)$.

Now we claim that the definition in terms of the map $\chi(x,y)$ is equivalent to the definition in terms of these two curves. We can use the first formulation of χ to prove this:

$$\chi(x_1, y_1, y_2) = y_1 \begin{bmatrix} 1-x_1 \\ (1-x_1^3)/3 \\ (1-x_1^6)/6 \end{bmatrix} + y_2 \begin{bmatrix} x_1 \\ x_1^3/3 \\ x_1^6/6 \end{bmatrix}$$

$$= y_1 \begin{bmatrix} 1 \\ \frac{1}{3} \\ \frac{1}{6} \end{bmatrix} + (y_2 - y_1) \begin{bmatrix} x_1 \\ \frac{1}{3}x_1^3 \\ \frac{1}{6}x_1^6 \end{bmatrix}$$

$$= (y_1 - y_2) \begin{bmatrix} 1-x_1 \\ \frac{1}{3}(1-x_1^3) \\ \frac{1}{6}(1-x_1^6) \end{bmatrix} + y_2 \begin{bmatrix} 1 \\ \frac{1}{3} \\ \frac{1}{6} \end{bmatrix}.$$

When $y_2 = y_1$, the image is the point $(2/3, 2/9, 1/9) = X$ as stated. When $y_2 > y_1$, we can use the left-hand expression to view the image as a point on the line segment between $\mathcal{C}_{4,1}$ and the curve $(t, t^3/3, t^6/6)$. When $y_2 < y_1$, the right-hand expression can be used to write the image as a point on the line segment between X and $(1-t, (1-t^3)/3, (1-t^6)/6)$. This means that the image of χ is contained in the regions and point specified.

To show that the reverse inclusion holds, we fix a point P in the interior of the convex hull of γ_1. By convexity, the line segment from X to P is contained in the region; continue in the direction $P - X$ until the line intersects the curve. This must occur because all points in the region are further from the bounding line than X. The point of intersection q is specified as $q = \gamma_1(\tau)$ for some $\tau \in (0,1)$. By convexity, there exists some ρ such that $\rho\ \mathcal{C}_{4,1} + (1-\rho)q = P$. Fixing $x_1 = \tau$, $y_1 = \rho$ and $y_2 = 1$, shows that P is in the image of χ. The same argument holds with slight variation for γ_2.

Proof of Proposition 4

The strategy to prove the equality of $\mathcal{C}_{4,3}$ and the cone over $\{t, t^3, t^6\}$ comes in two steps:

1. Show that the columns of $M_1(4,k)$ are always contained in the region R whose boundary is $\gamma_1 \cup \gamma_2 \cup \gamma_3$.

2. Divide the convex hull of R into two regions and show that each of these regions are included in $\widehat{\mathcal{C}}_{4,3}$.

First we demonstrate that the regions maps precisely into R. We have already shown in the main text of the document that the boundaries of $(0,1) \times (0,1)$ map to the boundaries of R under the mapping defined by $(x_1, x_2) \mapsto (x_1(1-x_2), x_1^3(1-x_2^3)/3, x_1^3(1-x_2^3)/6) \times 1/S$, where S is the sum of the coordinates. We compute the Jacobian of this map explicitly in `Macaulay2` (Grayson and Stillman 2002). The result is:

$$-\frac{1}{6S^3} x_1^9 (x_2 - 1)^4 (x_2^2 + x_2 + 1)(x_2^2 + 3x_2 + 1).$$

Plainly, this is nowhere zero in our domain. The inverse function theorem then implies that the interior is contained in the image of the boundaries. This accomplishes Step 1 of our proof.

For Step 2, we divide the image into two regions:

1. The triangle defined by vertices $(1,0,0)$, $(2/3,2/9,1/9)$ and $(1/3,1/3,1/3)$, including the two edges $[(1/3,1/3,1/3),(2/3,2/9,1/9)]$ and $[(2/3,2/9,1/9),(1,0,0)]$.

2. The remainder of the convex hull of R – explicitly, the interior of the region bounded by γ_3 and the line segment $[(1/3,1/3,1/3),(1,0,0)]$.

To show that the triangle is included, let $x_2 = \epsilon \approx 0$, and let x_1 vary. Then the third column sits arbitrarily close to $(1,0,0)$ and the first column traces out γ_2. Set $y_2 \approx 0$ and toggle y_1 and y_3, to obtain the full span, including the interior of the triangle, and the line segment $[(1/3,1/3,1/3),(2/3,2/9,1/9)]$. Set $x_1 = 1-\epsilon$, and the first column sits at $(1/3,1/3,1/3)$ while the third column traces out γ_1. This catches the missing line segment.

For the remainder of the convex hull, fix a point P in this region. This point lies on a line segment between $(2/3,2/9,1/9)$ and some point Q in γ_3. Suppose it is equal to $\rho \cdot (2/3,2/9,1/9) + (1-\rho)\cdot Q$. Set $x_2 = 1-\epsilon \approx 1$. We can choose ϵ and x_1 so that the second column is arbitrarily close to P. Furthermore, observe that the first column is approximately equal to the point on γ_2 corresponding to x_1 and the third column is approximately the point on γ_1 corresponding to x_1. Choosing $y_1 = y_3 = \rho$ and $y_2 = 1-\rho$ points us to

$$\rho \cdot \left(\begin{pmatrix} | \\ \gamma_1(x_1) \\ | \end{pmatrix} + \begin{pmatrix} | \\ \gamma_2(x_1) \\ | \end{pmatrix} \right) + (1-\rho)\cdot \begin{pmatrix} | \\ \gamma_3(x_1) \\ | \end{pmatrix}$$

$$= \rho \cdot \begin{pmatrix} 2/3 \\ 2/9 \\ 1/9 \end{pmatrix} + (1-\rho)\cdot Q \;=\; P.$$

Proof of Proposition 5

This is a direct application of the linear map W_4, computed as in Polanski and Kimmel (2003):

$$W_4 = \begin{pmatrix} 6/5 & 2 & 4/5 \\ 6/5 & 0 & -6/5 \\ 6/5 & -2 & 4/5 \end{pmatrix}.$$

Proof of Proposition 6

In order to prove the result about dimension, we show that $\mathcal{C}_{n,k}$ is a relatively open subset of a certain algebraic variety. Because the relevant operations are native to projective geometry, we transport our objects of interest in the obvious way to projective space. The same scaling properties that allow us to focus on the simplex also lead to good behavior in projective space.

Lemma 9. *For $k \geq 2$, the Zariski closure of $\mathcal{C}_{n,k}$ is the affine cone over $\mathcal{J}(\sigma_{k-2}(C_n, p_n))$, where:*

1. *the symbol C_n denotes the projective curve defined by mapping $[s:t]$ to*

$$C_n = \left[\binom{2}{2}^{-1} s^{\binom{n}{2}-\binom{2}{2}} t^{\binom{2}{2}} : \binom{3}{2}^{-1} s^{\binom{n}{2}-\binom{3}{2}} t^{\binom{3}{2}} : \cdots : \binom{n}{2}^{-1} t^{\binom{n}{2}} \right],$$

2. *the symbol p_n is the projective point $\left[1 : \frac{1}{3} : \frac{1}{6} : \cdots : \frac{1}{\binom{n}{2}}\right]$,*

3. *the operation $\mathcal{J}$ denotes the join of algebraic varieties, and*

4. *the operation $\sigma_i(\cdot)$ denotes the i-th secant variety. Following Harris (2013), the i-th secant variety is the union of i-dimensional planes generated by $i+1$ points in the variety.*

Proof of Lemma 9. The variety $\mathcal{J}(\sigma_{k-2}(C_n), p_n))$ is the image of the following map:

$$\psi(\vec{s},\vec{t},\vec{\lambda}) = \begin{pmatrix} 1 & s_1^{\binom{n}{2}-1} t_1 & \cdots & s_{k-1}^{\binom{n}{2}-1} t_{k-1} \\ \frac{1}{3} & \frac{1}{3} s_1^{\binom{n}{2}-3} t_1^3 & \cdots & \frac{1}{3} s_{k-1}^{\binom{n}{2}-3} t_{k-1}^3 \\ \vdots & \vdots & \ddots & \vdots \\ \frac{1}{\binom{n}{2}} & \frac{1}{\binom{n}{2}} t_1^{\binom{n}{2}} & \cdots & \frac{1}{\binom{n}{2}} t_{k-1}^{\binom{n}{2}} \end{pmatrix} \begin{pmatrix} \lambda_0 \\ \lambda_1 \\ \vdots \\ \lambda_{k-1} \end{pmatrix},$$

where s_i and t_i are not simultaneously zero, and λ is unrestricted.

Define the map $\phi : \mathbb{R}^{2k-1} \to (\mathbb{P}^1)^{k-1} \times \mathbb{R}^k$ sending $(x_1,\ldots,x_{k-1},y_1,\ldots,y_k)$ to

$$\left([1:x_1],[1:x_1x_2],\ldots,\left[1:\prod_{i=1}^{k-1} x_i\right], y_1, y_1+y_2,\ldots,\sum_{i=1}^{k} y_i \right).$$

We can recast the expression in (3) as the composition $\psi \circ \phi$. Based on this formulation, the set $\mathcal{C}_{n,k}$ is clearly contained in $\mathcal{J}(\sigma_{k-2}(C_n),p)$. To demonstrate the equality of the Zariski closures, we only need to show that the dimensions match and that the variety is irreducible. Both joins and secants have the property that irreducible inputs yield irreducible outputs, so the variety of interest is irreducible. The image of ϕ is open in $(\mathbb{P}^1)^{k-1} \times \mathbb{P}^{k-2}$, and the map ψ has deficient rank on a set of positive codimension. Therefore, the composition of $\psi \circ \phi$ has full dimension. This proves the Lemma. ☐

The i-th secant variety of an irreducible nondegenerate curve in $\mathbb{P}^n$ has projective dimension given by $\min(2i+1,n)$ (Harris 2013, Exercise 16.16). The curve C_n is a toric transformation of a coordinate projection of the rational normal curve. The rational normal curve is nondegenerate, and both of these operations preserve that property. This means our secant variety has projective dimension $\min(2(k-2)+1,n-2) = \min(2k-3,n-2)$. The join with a point adds 1 to the dimension of the variety, while the operation of passing to the affine cone adds 1 to the dimension of the variety *and* the ambient space. However, normalizing to the $(n-2)$-simplex subtracts 1 from both variety and ambient space again. This means that $\dim \widehat{\mathcal{C}}_{n,k} = \min(2k-2,n-2)$, assuming that $k \geq 2$.

Proof of upper bound in Theorem 8

Suppose a point $\mathbf{c}$ is in $\mathcal{C}_{n,q}$. By definition, this implies that there is a point $(x_1, \ldots, x_{q-1}, y_1, \ldots, y_q)$ such that (2) yields

$$\begin{bmatrix} 1-x_1 & x_1(1-x_2) & \cdots & \prod_{i=1}^{q-1} x_i \\ \frac{1}{3}(1-x_1^3) & \frac{1}{3}x_1^3(1-x_2^3) & \cdots & \frac{1}{3}\prod_{i=1}^{q-1} x_i^3 \\ \vdots & \vdots & \ddots & \vdots \\ \frac{1}{\binom{n}{2}}(1-x_1^{\binom{n}{2}}) & \frac{1}{\binom{n}{2}}x_1^{\binom{n}{2}}(1-x_2^{\binom{n}{2}}) & \cdots & \frac{1}{\binom{n}{2}}\prod_{i=1}^{q-1} x_i^{\binom{n}{2}} \end{bmatrix} \begin{bmatrix} y_1 \\ \vdots \\ y_q \end{bmatrix} = \begin{bmatrix} c_2 \\ \vdots \\ c_n \end{bmatrix}.$$

Since the point $\mathbf{c}$ is in the cone over the q columns of the matrix, Carathéodory's Theorem implies that it is also in the cone over some $n-1$ of the columns. Therefore we can replace the vector $y_1, \ldots, y_q$ with $y'_1, \ldots, y'_q$ so that all but $n-1$ (or fewer) are zero.

Passing to the expression in (3), this gives us:

$$\begin{bmatrix} 1 & x_1 & \cdots & \prod_{i=1}^{q-1} x_i \\ \frac{1}{3} & \frac{1}{3}x_1^3 & \cdots & \frac{1}{3}\prod_{i=1}^{q-1} x_i^3 \\ \vdots & \vdots & \ddots & \vdots \\ \frac{1}{\binom{n}{2}} & \frac{1}{\binom{n}{2}}x_1^{\binom{n}{2}} & \cdots & \frac{1}{\binom{n}{2}}\prod_{i=1}^{q-1} x_i^{\binom{n}{2}} \end{bmatrix} \begin{bmatrix} y'_1 \\ y'_2 - y'_1 \\ \vdots \\ y'_q - y'_{q-1} \end{bmatrix} = \begin{bmatrix} c_2 \\ \vdots \\ c_n \end{bmatrix}.$$

Since at most $n-1$ of the y'_i are nonzero, at most $2n-2$ of the entries of the vector at right are nonzero. We delete the columns of the X matrix corresponding to zero entries except the first column. A new sequence $(x'_1, \ldots, x'_{2n-2})$ may then be obtained from the ratio between the first entries in adjacent columns. The new sequence $y''_1, \ldots, y''_{2n-1}$ is obtained by taking the sequence of partial sums of the vector.

Proof of non-convexity in Theorem 8

To prove this final result, we combine two properties already proven:

1. The manifold $\mathcal{C}_{n,k}$ is a proper subset of $\mathcal{C}_{n,k+1}$ for all $k < \kappa_n$ (from Proposition 6).

2. The manifold $\mathcal{C}_{n,\kappa_n}$ is contained in the convex hull of $\mathcal{C}_{n,2}$. (This follows from Equation 2.)

Since $\mathcal{C}_{n,k}$ contains $\mathcal{C}_{n,2}$ and is properly contained in the convex hull of $\mathcal{C}_{n,2}$, it cannot be convex.

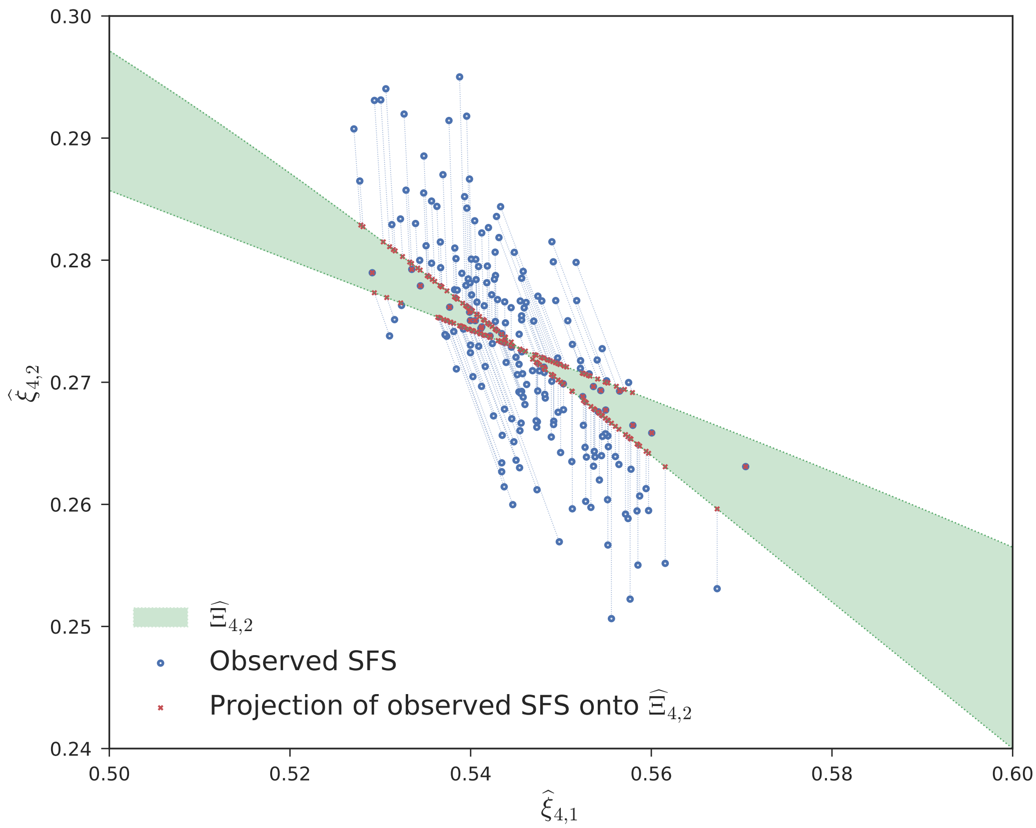

Figure 1 The green region, denoted $\widehat{\Xi}_{4,2}$, represents the set of expected SFS for two-epoch piecewise-constant demographies for sample size $n = 4$. Each blue circle is the observed SFS simulated using `msprime` (Kelleher *et al.* 2016) under a constant population size coalescent with recombination using realistic mutation and recombination rates of 10^{-8} mutations and 2.2×10^{-8} crossovers per basepair per generation per haploid. Each sequence has 1000 unlinked loci of length 10 kb each, resulting in an average of 7,300 segregating sites. The red crosses are the expected SFS inferred for these simulated SFS using `fastNeutrino` (Bhaskar *et al.* 2015); the dotted blue lines show the correspondence between the observed SFS and their projections onto $\widehat{\Xi}_{4,2}$. For observed SFS lying in the interior of $\widehat{\Xi}_{4,2}$, the observed SFS and their projections coincide, while the observed SFS lying outside $\widehat{\Xi}_{4,2}$ project onto the boundaries of one of the two convex regions that form $\widehat{\Xi}_{4,2}$.

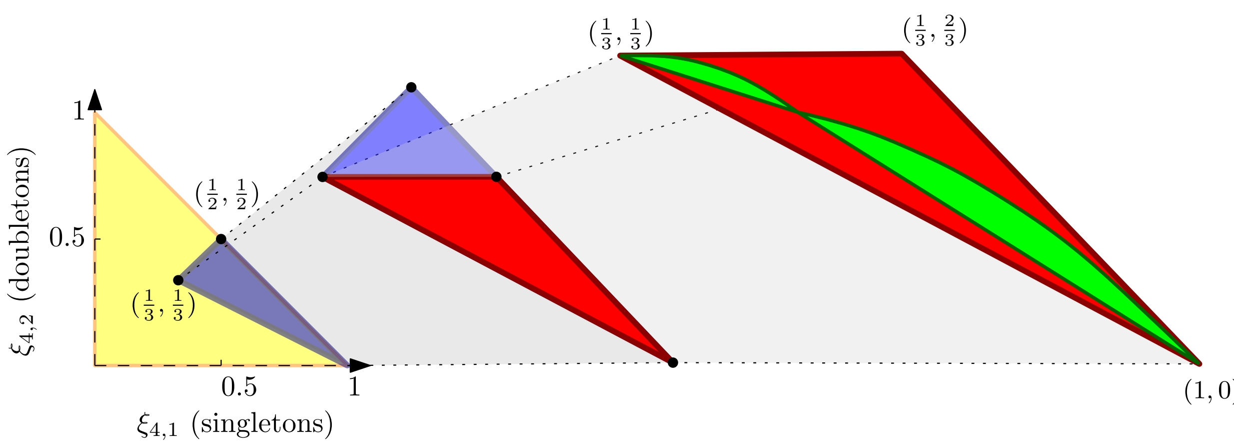

Figure 2 Eliminating candidate normalized SFS vectors for $\widehat{\Xi}_{4,2}$. This image considers candidate vectors and eliminates them for different reasons. *A priori*, any vector adding up to 1 is a possible SFS. This is represented by the yellow triangle whose third coordinate (not shown) is simply one minus the sum of the other two. Sargsyan and Wakeley (2008) showed that the SFS is non-decreasing, ruling out any vectors outside the blue triangle. Furthermore, they showed that the SFS is convex, therefore $\xi_{4,2} \leq \frac{1}{2}(\xi_{4,1} + \xi_{4,3})$, ruling out anything outside the red triangle. Finally, our algebraic analysis of the expected SFS for a piecewise-constant demography with two epochs rules out vectors outside the green region at right.

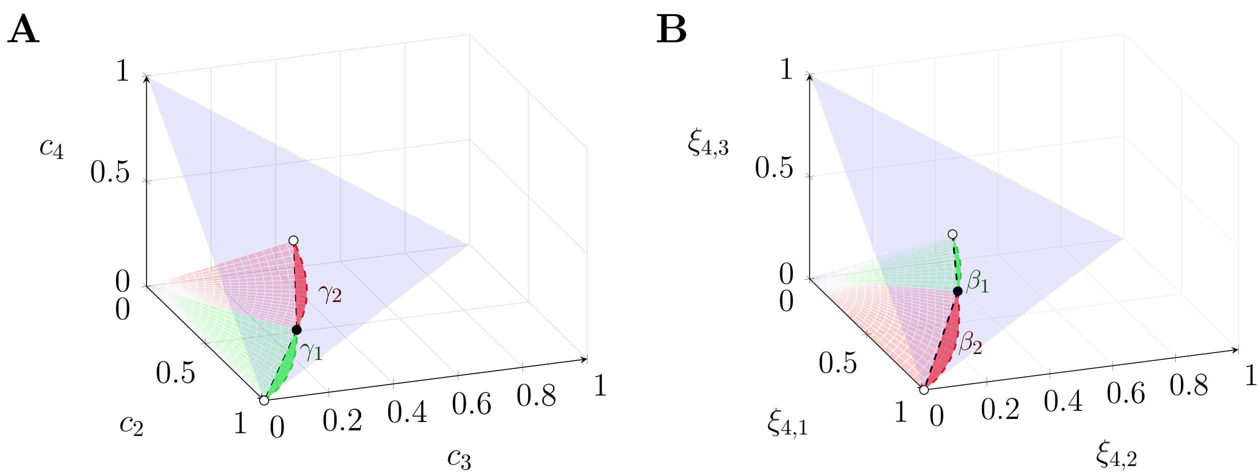

Figure 3 Coalescence and SFS manifolds for sample size 4 and 2 population epochs. **A.** The coalescence manifold $\mathcal{C}_{4,2}$ is the union of red and green cones. The 2-simplex, shaded in blue, intersects $\mathcal{C}_{4,2}$ in the normalized coalescence manifold $\widehat{\mathcal{C}}_{4,2}$. The green region corresponds to recent-small, ancient-large demographies; the red region to recent-large, ancient-small demographies. **B.** The SFS manifold $\Xi_{4,2}$ is the union of red and green cones. The 2-simplex intersects $\Xi_{4,2}$ in the normalized SFS manifold $\widehat{\Xi}_{4,2}$. Here, too, the green region corresponds to small-then-large demographies; the red region to large-then-small demographies. As mentioned earlier, $\Xi_{4,2}$ is obtained from $\mathcal{C}_{4,2}$ by a linear transformation.

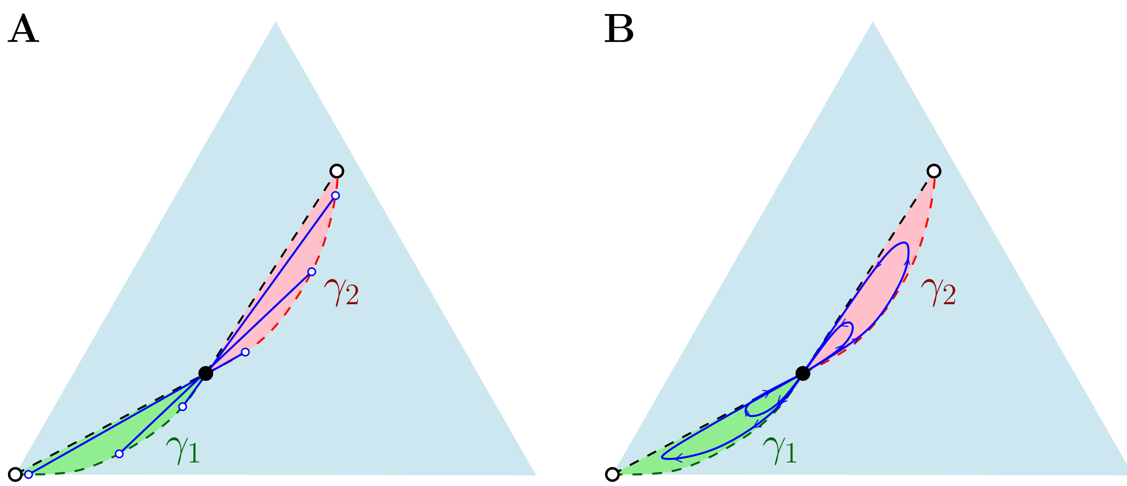

Figure 4 Fixed-time and fixed-size contours in $\widehat{\mathcal{C}}_{4,2}$. A. The blue line segments correspond to the image of $\chi_{4,2}(x^*, \vec{y})$ where x^* is a constant fixing the break-point between the two demographies. The other input $\vec{y} = (y_1, y_2)$ varies over all positive vectors, though scaled $\vec{y}$ vectors point to the same normalized value. As $y_1/y_2 \to 0$, the image approaches γ_1 and as $y_2/y_1 \to 0$, the image approaches γ_2. **B.** The blue curves correspond to the image of $\chi_{4,2}(x, \vec{y}^*)$ where $\vec{y}^*$ is a fixed vector indicating the population values and x takes all values in $(0, 1)$. The endpoints 0 and 1 correspond to breakpoints at ∞ and 0 respectively. For $y_1^* < y_2^*$, x traces a loop in the green region; for $y_1^* > y_2^*$, x traces a loop in the red region.

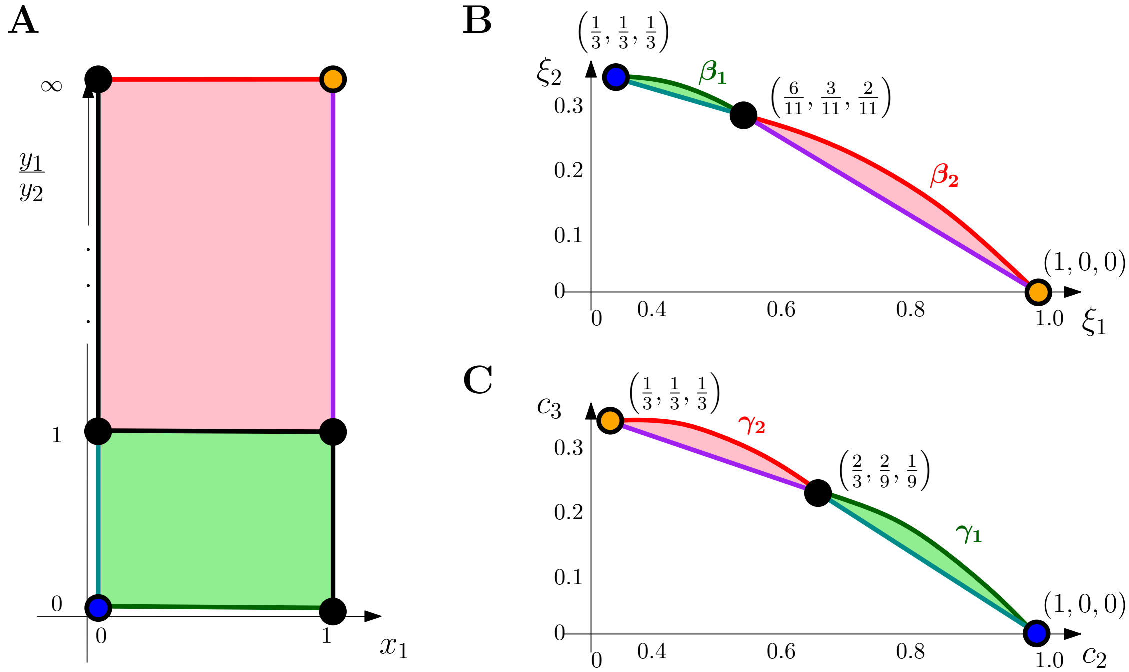

Figure 5 Pairing the boundaries of demography space and $\widehat{\mathcal{C}}_{4,2}$. **A.** The domain of $\chi_{4,2}$. Note that for fixed y_1/y_2, the normalized coalescence vector is the same. **B.** The normalized SFS manifold $\widehat{\Xi}_{4,2}$ projected onto its first two coordinates. **C.** The normalized coalescence manifold $\widehat{\mathcal{C}}_{4,2}$ projected onto its first two coordinates. The red square at left corresponding to $y_1 > y_2$ maps to the red regions at right; the green square at left corresponding to $y_2 < y_1$ maps to the green regions at right. The black line segments on left (corresponding to $y_1/y_2 = 1$; $y_2 < y_1$ and $x_1 = 0$ (equivalently $t_1 = \infty$); $y_2 > y_1$ and $x_1 = 1$ (equivalently $t_1 = 0$)) all map to the central black points on right, since they each mimic a constant demography. The green line corresponding to $y_1 = 0$ maps to the curve β_1 in $\widehat{\Xi}_{4,2}$ and the curve γ_1 in $\widehat{\mathcal{C}}_{4,2}$; the red line corresponding to $y_2 = 0$ maps to the curve β_2 in $\widehat{\Xi}_{4,2}$ and the curve γ_2 in $\widehat{\mathcal{C}}_{4,2}$. The orange point ($x_1 = 1, y_2 = 0$) maps to $(\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$ in $\widehat{\mathcal{C}}_{4,2}$ and maps to $(1, 0, 0)$ in $\widehat{\Xi}_{4,2}$. The blue point ($x_1 = 0, y_1 = 0$) maps to $(1, 0, 0)$ in $\widehat{\mathcal{C}}_{4,2}$ and $(\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$ in $\widehat{\Xi}_{4,2}$. The remaining aqua and violet segments map to the segments of the same color.

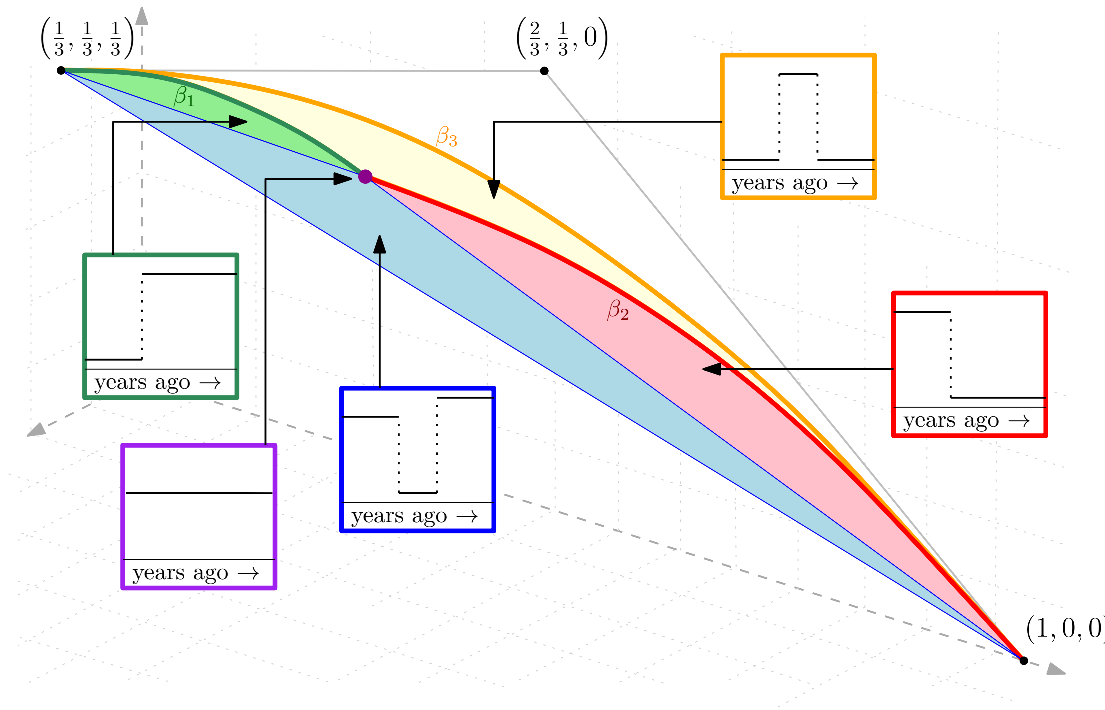

Figure 6 Regions of $\widehat{\Xi}_{4,3}$ and sample demographies. The image depicts $\Xi_{4,3}$ partitioned into different colored regions. The purple point in the center is the SFS corresponding to the constant demography. The green region contains SFS corresponding to recent-small, ancient-large demographies. The red region corresponds to recent-large, ancient-small demographies. The orange region contains SFS corresponding to three-epoch demographies with a boom in the second epoch. The blue region contains SFS corresponding to three-epoch demographies with a bottleneck in the second epoch. These are not the unique demographies mapping to each region of $\Xi_{4,3}$, but they depict, in some sense, the simplest demographies yielding those SFS.

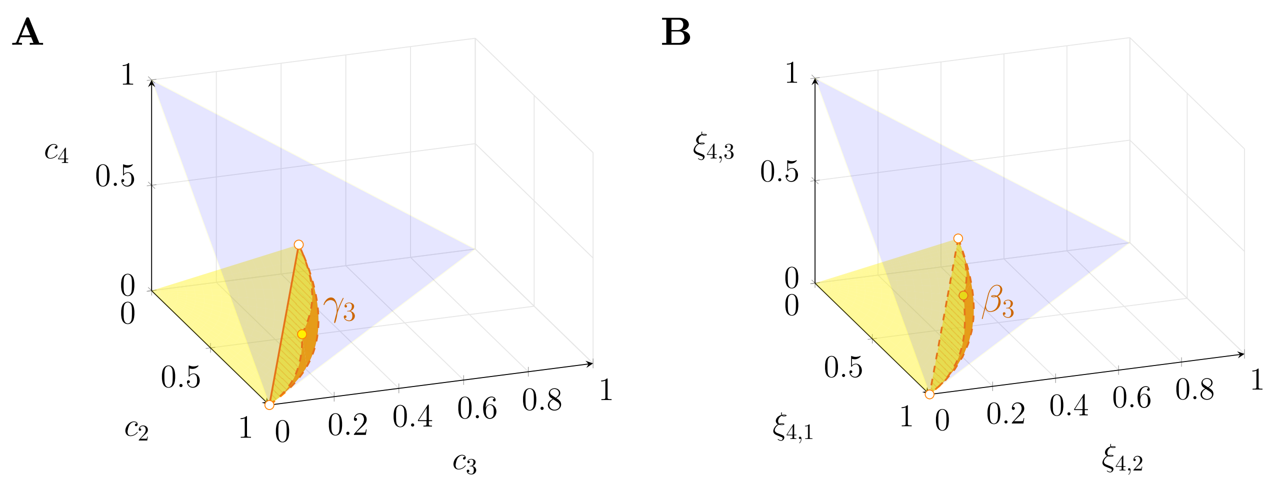

Figure 7 Coalescence and SFS manifolds for sample size 4 and 3 population epochs. **A.** The coalescence manifold $\mathcal{C}_{4,3}$ is the entire yellow and orange region. The 2-simplex, shaded in blue, intersects $\mathcal{C}_{4,3}$ in the normalized coalescence manifold $\widehat{\mathcal{C}}_{4,3}$. The orange region of $\widehat{\mathcal{C}}_{4,3}$, bounded by γ_1, γ_2, and γ_3, is the image of the surface described by the columns of $M_1(4,3)$, while the yellow region adds in vectors gained by using convex combinations. **B.** The SFS manifold $\Xi_{4,3}$ is the entire yellow and orange region. The 2-simplex intersects $\Xi_{4,3}$ in the normalized SFS manifold $\widehat{\Xi}_{4,3}$. The SFS manifold $\Xi_{4,3}$ is obtained from $\mathcal{C}_{4,3}$ by a linear transformation. The orange region of $\widehat{\Xi}_{4,3}$, bounded by β_1, β_2, and β_3, is the image of the surface described by the columns of $M_1(4,3)$, while the yellow region adds in vectors gained by using linear combinations.